Sparsity Ranked Lasso with Partial Autocorrelation and Exogenous Variables - SRLPAx

Federico Deotto, Mathias Steilen, Olivier Zehnder, Tito Quadri

The 'SRLPAx' model [1] was designed for use with time series that have a complex seasonal component and exogenous variables. The classical approach attempts to model these types of time series using a Seasonal Autoregressive Model (SAR) with local and seasonal components, denoted by ϕ and θ respectively. The model can be expressed as:

$$Y_t = \beta_0 + \sum_{j=1}^{p} \phi_j Y_{t-j} + \sum_{j=1}^{P} \theta_j Y_{t-jm} + \epsilon_t$$

where the innovations ϵ_t are assumed to be normally distributed around zero with constant variance. There are also other versions of this model that integrate a moving average component, such as the ARMA. Usually, to model the seasonal component of a time series, new terms are introduced into the model, such as nearby seasonal lags, rewriting the model as:

$$Y_t = \beta_0 + \sum_{j=1}^{p} \phi_j Y_{t-j} + \sum_{j=1}^{P} \theta_j^T Y_{(t-jm-c):(t-jm+c)} + \epsilon_t$$

where $Y_{a:b}^T = (Y_a, Y_{a+1}, \ldots, Y_{b-1}, Y_b)$ indicates the observations ranging from time a to time b, and θ_j refers to the vector of seasonal coefficients on lags centered at $t - jm$. The rationale for including these terms is that it might be uncertain which lags near $t - jm$ are relevant; this method encompasses all lags around $t - jm$ within c that could play a role in modeling Y_t.

The formula of the 'SRLPAx' model, used in this project, is:

$$Y_t = \beta_0 + \sum_{k=1}^{d} \beta_k X_{t,k} + \sum_{j=1}^{p^*} \phi_j Y_{t-j} + \epsilon_t$$

and its loss is:

$$\sum_{t=p^*+1}^{n} \left\{ \|y_t - \beta_0 - \sum_{k=1}^{d} \beta_k x_{t,k} - \sum_{j=1}^{p^*} \phi_j y_{t-j}\|_2^2 \right\} + \lambda \left\{ \sum_{j=1}^{p^*} w_j |\phi_j| + \sum_{k=1}^{d} w_k' |\beta_k| \right\}$$

where $x_{t,k}$ is the k-th observed exogenous variable associated with the variable Y_t. The exogenous variables, although indexed by t, can be observed in a different preceding time period. The β_k represents the coefficient that expresses the effect of the exogenous variables X_k on the response Y. In particular, β_0 is the intercept of the model.

The autoregressive part of this model corresponds to an AR(p^*) model where $p^* = P \cdot m + c$.

The weights w_j that regulate the penalizations of the autoregressive component of the response are informed through the partial autocorrelation function (PACF):

$$w_j = \frac{1}{|\hat{\phi}_j|^{\gamma}}$$

where $\hat{\phi}_j$ is the estimated PACF of the variable Y_t with respect to the j-th lag. Notice that when γ is zero, the penalization is the same as in the traditional L1 penalization since the weights w_j are equal to one.

The weights w'_k associated with the exogenous variables are obtained as the slope of the regression line between the variable Y_t and the exogenous variable $X_{t,k}$. The downside of including a penalization for the exogenous variables in the model is the cost of not being able to perform inference on their estimated coefficients in a classical statistical sense without correcting for post-selection inference or accounting for bias due to the shrinkage of the coefficients. What we can do is simply observe those coefficients that have been selected by the model, disregarding the null ones, and consider as more relevant in describing the phenomena those variables associated with estimated coefficients having larger absolute values.

Also, this model considers a normally distributed innovation ϵ_t centered around zero with constant variance.

The SRLPAx should be used instead of an ARMA model when there are many exogenous variables X, the number of observations n is large, or the autoregressive component p is large, since the normal procedures to find the parameters of the model are greedy. The SRLPAx allows estimating the values of the parameters and simultaneously selecting the most appropriate values for p and P by using a weighted L1 penalization.

A final remark about an important assumption regarding the effect of the exogenous variables on the response variable, namely, that their effect is constant over time. In the context of our project, this implies that the impact of fossil fuels on electricity production should remain consistent throughout the time window under consideration. However, this assumption can be debatable, especially if we consider a period that begins before and ends after the Ukrainian war. Moreover, the electricity market is a physical market, which differentiates it from a financial market. However, it still shares similarities with financial markets, particularly in the presence of extreme values that are difficult to model using a normal distribution, which has thin tails.

References

[1] Peterson R, Cavanaugh J. Fast, effective, and coherent time series modelling using the sparsity-ranked lasso. *Statistical Modelling*. 2024;0(0). doi:10.1177/1471082X231225307.